

DOCUMENT RESUME

ED 143 688

TM 006 444

AUTHOR Haladyna, Tom
TITLE Measuring Performance: Teacher-Made Tests.
INSTITUTION Oregon State Dept. of Education, Salem.
PUB DATE 77
NOTE 58p.
EDRS PRICE MF-\$0.83 HC-\$3.50 Plus Postage.
DESCRIPTORS Achievement Tests; Attitude Tests; Essay Tests; Evaluation Criteria; Guidelines; *Guides; Multiple Choice Tests; Rating Scales; Student Evaluation; *Teacher Developed Materials; *Test Construction; Test Interpretation; Test Items; Test Reliability; *Tests; Test Validity

ABSTRACT

Among the new testing developments are the use of objectives or goals in instruction, competency based approaches to instruction, criterion referenced testing, and performance oriented testing. These new approaches often emphasize individualized learning; each student's progress is individually monitored by comparison with clear statements of what students are expected to learn. The careful monitoring of individual student progress required in such a performance based system has created the need for a testing technology which differs in many respects from typical practice. These guidelines are intended to help teachers develop testing skills which meet the demands of a competency based approach to instruction. This guide includes information on fundamental concepts in testing, construction of classroom tests and measurement of attitudes. An annotated bibliography of recent and significant contributions to testing technology, a brief glossary of testing terms, and a list of recommended sources for achievement test items are appended.

(Author/MV)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED143688

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

MEASURING PERFORMANCE: Teacher-Made Tests

REPRODUCED BY PERMISSION FROM
THE NATIONAL INSTITUTE OF EDUCATION

*Oregon State Dept.
of Education*

THIS DOCUMENT HAS BEEN REPRODUCED
EXACTLY AS RECEIVED FROM THE
PERSON OR ORGANIZATION ORIGINATING IT.

TM006 444

MEASURING PERFORMANCE:
TEACHER-MADE TESTS

Verne A. Duncan
State Superintendent of
Public Instruction

Oregon Department of Education
942 Lancaster Drive NE
Salem, Oregon 97310

1977

STATEMENT OF ASSURANCE

Oregon Department of Education

It is the policy of the Oregon Department of Education that no person be subjected to discrimination on the basis of race, national origin, religion, sex, age, handicap, or marital status in any program, service, or activity for which the Oregon Department of Education is responsible. The Department will comply with the requirements of state and federal law concerning nondiscrimination and will strive by its actions to enhance the dignity and worth of all persons.

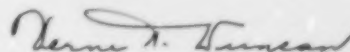
45231197710500

FOREWORD

In June 1976, the State Board of Education adopted revised minimum standards for Oregon public schools. A response to citizen concerns regarding what is, in fact, expected of schools, the standards call for a system of goal-based planning, which includes testing and assessment procedures.

The Department of Education is committed to helping districts implement the standards. Current and anticipated problems are being identified, priorities set, and resources allocated.

One priority area centers on the assessment requirements found in the standards. Measuring Performance: Teacher-Made Tests is one of a series of publications dealing with assessment. It focuses on helping teachers improve their techniques for building tests to assess growth in student achievement. It is my hope that this and other publications in the assessment series prove useful in implementing district practices that will meet the intent of the planning and assessment requirements. For further information, contact the Department's Director of Evaluation and Assessment, 942 Lancaster Drive NE, Salem 97310, telephone 378-3074.



Verne A. Duncan
State Superintendent of
Public Instruction

ACKNOWLEDGEMENTS

Extensive field testing of this material has been conducted with classroom teachers and administrators across the State of Oregon. To those of you who have read and reacted to the materials in whole or in part--thanks for helping to make this document as easy to read and understand as possible.

Special recognition is due Tom Haladyna, Associate Research Professor, Teaching Research Division, State System of Higher Education. Doctor Haladyna was principal author of the initial draft of the document under a contract with the Planning, Evaluation, and Assessment Program, Oregon Department of Education.

TABLE OF CONTENTS

	<u>Page</u>
Foreword	iii
Acknowledgements	v
Table of Contents	vii
Introduction	1
 I. FUNDAMENTAL CONCEPTS IN TESTING	 3
Interpreting the Measurement	3
Criteria for Test Quality	3
Matching Items to Instructional Intent	4
Types of School Outcomes	6
Self-Quiz	7
 II. CLASSROOM TEST CONSTRUCTION: Selected Response Test Items . .	 9
Multiple Choice Items	10
Selecting Items	11
Writing Good Multiple Choice Items	11
Other Deficiencies in Item Writing	12
True-False Tests	14
Matching Items	16
Item Forms	16
Self-Quiz	17
 III. CLASSROOM TEST CONSTRUCTION: Constructed Response Items . . .	 19
When Should We Use the Essay Test?	19
Writing a Short Answer Question	20
Test Length and Format	21
Scoring the Response	21
Self-Quiz	22
 IV. CLASSROOM TEST CONSTRUCTION: Measuring Skills	 25
Observation	26
Rating Scales	26
Factors Affecting the Reliability of Ratings	31
Checklists	31
Self-Quiz	33
 V. MEASURING ATTITUDES.	 35
Approaches to Measuring Attitudes	35
Scoring and Analyzing Results	38
Observation Methods	39

	<u>Page</u>
V. MEASURING ATTITUDE (Continued)	
General Observations	40
Critical Incidents	41
Self-Quiz	43
APPENDIX A Annotated Bibliography	45
APPENDIX B Glossary of Terms	49
APPENDIX C Sources of Achievement Test Items	51

INTRODUCTION

Recent developments in teaching and testing may alter dramatically that which we have called "traditional education." To some extent, these innovations are responses to what appears to be a decline in school achievement at all grade levels. Among these new developments are: the use of objectives or goals in instruction, competency-based approaches to instruction, criterion-referenced testing, and performance-oriented testing.

The new approaches often emphasize individualized learning: each student's progress is individually monitored by comparison with clear statements of what students are expected to learn. Since individualized instruction must be flexible and responsive to individual needs, progress should not be determined by how one student compares with another. Instead, the growth of each student should be determined by comparison against previously determined standards.

The careful monitoring of individual student progress required in such a system has created the need for a testing technology which differs in many respects from typical practice. These guidelines are intended to help teachers develop testing skills which meet the demands of a competency-based approach to instruction.

CHAPTER I: FUNDAMENTAL CONCEPTS IN TESTING

Any judgment about whether or not learning has occurred is an inference based on measurement. The more accurate the measurement, the more likely the inference will be correct. The need for information about how well children have learned leads to the construction of exercises (items) which can be used to assigned numerical value to various levels of achievement. A collection of such exercises is usually called a test. A test should be uniform and standard in several ways; for example, when a test is administered to a group of students, all should receive the same items under the same conditions for about the same amount of time.

The sum of correct responses to the items on a test constitutes the score. Scores may be compared to one another or to some desirable standard, often called a "passing standard."

Interpreting the Measurement

Any measure (including test scores) must have a comparative basis for interpretation. The basis for test scores can be either standard-referenced or grouped-referenced. If there are 24 principles of science to learn in a unit, with two test items for each principle, a test score of 32 on a 48-item test could be reported either way. The standard-referenced score is about 67 percent; it might be inferred that the student understands about two-thirds or 67 percent of the principles of science in that unit, as judged by test results. If the acceptable standard is 75 percent of items of a given difficulty, then 67 percent is not an acceptable level of achievement. The standard could be set at 100 percent if it is desirable for students to respond correctly to all items of this difficulty.

Using the group-referenced method, a score of 32 may be the highest in the class. On a larger scale, it may be higher than the scores of 60 percent of a cross-section of other sixth grade students. This approach is useful in comparing students to a previously selected reference group, reassigning high or low scoring students to different groups, or helping a student select a course of study or investigate possible future occupations.

Criteria for Test Quality

There are three major criteria by which the quality of any measure may be judged: accuracy, precision, and efficiency.

In educational measurement, the accuracy with which a certain characteristic is represented by scores on a given test instrument is called "validity." Validity relates to the content of the test. It represents the degree to which a test actually measures what it purports to measure--usually a change in behavior which is inferred to have resulted from a particular set of teaching/learning activities.

Precision is concerned with the amount of error associated with any measurement. Measurement errors usually occur randomly, but are distributed normally. Therefore, any particular measurement contains some error--large or small, positive or negative. The frustrating fact is that the exact size of the error and the direction in which it occurs, positive or negative, are unknown for any particular measurement.

The degree of precision, or consistency, of measurement using a test instrument is referred to as test reliability. Any individual test score is composed of what might be called a "true" score value plus or minus a margin of error. The more reliable a test is, the smaller the amount of measurement error that is associated with each individual score. The reliability of a test is especially important if scores are going to be used in a pass/no pass situation (e.g., a situation in which a score of 63 or higher is a passing score--62 or less is not).

In some test situations the stress of the measurement itself is a source of variation. This would be analogous to a situation in which attempts to measure the length of a table somehow actually caused changes in its length! If test scores are to be useful, they must be reliable across a variety of circumstances and from one instance to another. Reliability, or the consistency with which a trait can be represented by a test score, is essential to good educational measurement.

Efficiency involves the time consumed and money spent in planning, constructing, administering, scoring, and reporting test results. Efficiency is considered in terms of the teacher and the student.

Testing practices which lead to highly accurate and precise measures are generally inefficient. Test-makers usually try to achieve a balance with accuracy and precision on the one hand, and efficiency on the other.

Matching Items to Instructional Intent

The purpose of most teacher-made classroom tests is to measure changes in achievement level that are a presumed result of teaching. These types of tests are not usually employed to measure other related traits, such as aptitude for learning, motivation, and attitude toward school. (However, such traits are important. The measurement of attitudes is discussed further in Chapter V.) Since the test item is the basic unit of measurement, each item must be of high quality and must accurately represent the teacher's instructional intent. The objective, or instructional goal, is the most often advocated and best known device for connecting items with teaching intent. While there are several sets of rules for what constitutes a good instructional goal, any statement which communicates the instructional intent of the teacher can be used to link that intent with the test item. More specifically, an objective or goal can be any statement that describes what the student will be required or expected to do under certain conditions (learning outcomes). The following pages provide information about types of learning outcomes and describes techniques for writing outcome statements which satisfy measurement requirements. Four types of outcomes are presented in Table 2.

Table 2

Types of Outcomes and Related Measurement Techniques

DOMAIN	Cognitive		Affective	Psychomotor
OUTCOMES	Knowledge	Skills	Attitudes	Skills
ASPECTS	Recall Comprehension or Higher Level Behaviors ²	Performance Products		Performance
NATURE OF TRAIT	Inferred in a paper-and-pencil instrument	Inferred or Directly Observed	Inferred or Directly Observed	Directly Observed
TECHNIQUES	<u>Selected Response</u> 1. multiple choice 2. true-false 3. matching <u>Constructed Response</u> 1. completion 2. short answer essay 3. extended answer essay	<u>Observation</u> <u>Rating Scales</u> 1. numerical scales 2. graphic scales 3. descriptive scales <u>Checklists</u>	<u>Observation</u> <u>Rating Scales</u> 1. numerical scales 2. graphic scales 3. descriptive scales <u>Checklists</u> <u>Anecdotal Records</u>	<u>Observation</u> <u>Rating Scales</u> 1. numerical scales 2. graphic scales 3. descriptive scales <u>Checklists</u>

²Methods of measuring higher level behaviors are discussed in: Bloom (1956), Miller, Williams, and Haladyna (in press), or Sanders (1966); see Appendix A.

Types of School Outcomes

It is generally accepted that the mission of the public school is to help each student acquire the knowledge, skills, and attitudes necessary to function effectively in a variety of life roles. Knowledge and skills are usually associated with cognitive behavior. Psychomotor skills are related to movement, such as marching, but are often found combined with the others, as in the playing of a musical instrument. Attitudes are merely one aspect of affective behavior which may or may not be significantly related to cognitive behavior. There has been increasing interest in measuring student attitude as a result of the growing conviction that a student's attitude toward school and society may be as important as knowledge or skills acquired.

Knowledge. Knowledge is an attribute which can only be inferred from student behavior; it is not directly observable; it is intangible. The acquisition of knowledge is not usually measured through responses to paper-and-pencil tests, oral responses, or an observed physical response. Some paper-and-pencil instruments use selected response techniques, including multiple choice, true-false, or matching. The latter two are actually a variation of the first. Another technique is to ask students to construct their answers to questions rather than to select an appropriate response. Constructed response techniques include completion, short answer essay, and extended answer essay. While constructed response tests have some limitations, they can be useful in measuring school achievement. Selected response testing is discussed further in Chapter III and constructed response testing in Chapter IV.

Skills. Measuring skills typically involves the observation of a performance or product. The techniques most used for measuring skills are (1) observation, (2) rating scales (rendering judgments on some numerical scales), or (3) checklists (marking whether a sequence of behaviors or traits is present or absent). Techniques for measuring skills are presented in Chapter V.

Attitudes. The third type of outcome, attitudes, is in the affective domain. As shown in Table 2, the measurement of attitude involves many of the same techniques that were used to measure skills. There is one additional technique, however: anecdotal records. The measurement of attitudes is discussed in greater detail in Chapter VI.

A variety of techniques can be used to measure most outcomes; in some cases, however, the necessary technique is self-evident.

SELF-QUIZ

Match the techniques on the right with the statements of instructional intent on the left.

<u>OUTCOME STATEMENT</u>	<u>TECHNIQUES</u>
1. Given pictures of animals, the student will identify the type of invertebrate or vertebrates.	a. observation
2. Given sentences, the student will select the correct verb tense.	b. short answer essay
3. The student will describe the action of molecules in melting ice.	c. multiple choice
4. The student will determine the speed of a falling body when given the number of seconds the body has fallen and the law of falling bodies.	d. rating scales
5. The student will provide a rationale for flu shots for the aged.	e. checklists
6. The student will correctly reassemble a motor.	f. completion
7. Students will be able to complete a manual dexterity exercise in less than two minutes.	
8. Name and briefly describe, in writing, four aspects of effective communication.	
9. How effective is the use of lubrication in improving the mechanical efficiency of a machine?	
10. How well has the student used alliteration in his poem?	

OUR ANSWERS

1. C, multiple choice was selected because knowledge is being measured and it would be efficient to provide options for the student for each picture. A completion format would also be appropriate here.
2. C, again selecting the correct answer seems most appropriate for this knowledge outcome.
3. B, since description if required, the short answer essay seems best.
4. F, this is a higher level of knowledge requiring careful analysis and a constructed response.
5. B, the rationale is a knowledge outcome, best presented in a short answer essay.
6. E, implicit here is that a sequence of actions must occur and must be performed correctly in order for the motor to be correctly assembled.
7. A, this is a simple observation; the teacher is interested in a psychomotor outcome.
8. B, a measurement of knowledge which is most appropriately expressed in a short answer essay.
9. D, appears to ask for a rating of effectiveness, a performance-oriented outcome.
10. D, an estimate of how well the student's product exhibits a desired learning outcome.

CHAPTER II: CLASSROOM TEST CONSTRUCTION

Selected Response Test Items

Classroom achievement tests should measure learning in relation to teaching objectives, using the most comprehensive sampling possible. Selected response tests (i.e., including multiple choice, true-false, and matching) offer the greatest potential for sampling because more questions can be asked in a set length of time. A longer test (i.e., one with more items) is usually more accurate and precise. Thus the selected response technique is most often used by measurement experts and teachers.

Potential Advantages

- + Easy to score.
- + Can be objectively scored; that is, the answer is predetermined and any grader should arrive at the same test score.
- + Can be used to measure simple recall or higher level knowledge.
- + Generally requires less time per item on the part of the student; therefore, more questions can be answered and a greater sampling of the material can be achieved.

Potential Disadvantages

- Correct answers may sometimes be inferred from the item itself unless questions are carefully worded.
- Some items test recall of trivial knowledge.
- Time-consuming and difficult to construct accurate and precise items.
- Correct answers may be achieved by guessing.

These disadvantages are usually considered minor in significance. For example, guessing plays a small part in most tests. If there are four choices for each item in a well-constructed 100-item multiple choice test, a student lacking knowledge would be expected to score about 25 percent; on the average, students would guess correctly once in every four times. Test scores should be interpreted with this in mind. While the second and third disadvantages are more common, they can be overcome. (See Appendix A for titles of guides to writing multiple choice items for higher level knowledge.) While these tests are more difficult to construct, they are easier to score. Items that prove useful can be used year after year, provided the teaching objectives do not change.

Multiple Choice Items

Multiple choice items are usually written in one of two ways:

- | | | |
|--------------------|---|----------------|
| Mary had a little | → | item stem |
| a. too much to eat | → | foil |
| *b. lamb | → | correct answer |
| c. goat | → | foil |
| d. brother | → | foil |

- | | | |
|---------------------|---|----------------|
| What did Mary have? | → | item stem |
| a. too much to eat | → | foil |
| *b. a little lamb | → | correct answer |
| c. a little goat | → | foil |
| d. a tiny brother | → | foil |

Note the item stem. Wrong answers are foils (distractors). Collectively, the correct answer and the foils are called options. Well-written items have foils which are likely to be chosen by students who have not yet learned the objective which the item was designed to measure. That is, all options should be feasibly correct responses if the student does not know the material.

The first item example has an item stem which is part of a sentence and is completed with the options. The second item stem is a question, and the options answer the question. All options should be parallel in grammatical construction and length.

Both of the following items are examples of how not to construct a multiple choice question.

- Jack and Jill
- a. were two nice kids
 - b. boy and a girl
 - c. They were going down the hill
 - d. I don't know

This item illustrates incorrect grammatical construction. Only option "a" completes the sentence.

- Where did the cow jump?
- a. 7' 3-1/2" for a new world's record
 - b. over the moon
 - c. She jumped over the fence
 - d. wherever she wanted

Each option has a different grammatical construction.

Selecting Items

Many items already exist that could accurately measure the teacher's instructional intent. These can be collected, organized, and used to great advantage. It saves teacher time in constructing items and helps build effective and accurate tests. Appendix C lists some sources for items. The teacher might even ask students to help write items. Provide student the item stems and ask them to suggest options. Students can also be helpful critics as teachers rewrite to improve the quality of particular test items.

Writing Good Multiple Choice Items

Procedures:

1. Identify the concept or objective to be tested.
2. Write the item stem.
3. Write the correct answer.
4. Write the foils parallel in structure to the correct answer.

Item writing gets easier with practice. The following exercise illustrates some pitfalls that can be avoided.

EXERCISE

Before you take this test, be forewarned that you are expected to score 100 percent. Mark the letter corresponding to your answer in the space provided. Try to "figure out" the right answers using the clues given.

- ____ 1. How often were the seven Cities of Cibola discussed in early Spanish literature?
 - a. seldom
 - b. always
 - c. never
 - d. all of the time
- ____ 2. The beta-binomial distribution is a function of which of the following theoretical distributions?
 - a. bivariate normal
 - b. multivariate normal
 - c. poisson
 - d. beta
- ____ 3. Which Romantic poet is best known for his "amorous" activities?
 - a. O. J. Simpson
 - b. Byron
 - c. "the Fonz"
 - d. Henry Kissinger

- ____ 4. The thermotrople adjustment is
 a. sixteen
 b. $8 + 8$
 c. 2
 d. 4×4
- ____ 5. The most defensible reason for using caution in handling hydrochloric acid is the
 a. heat
 b. smell
 c. weight
 d. fact that when it comes into contact with the skin it will burn you.

OUR ANSWERS

D *S C *A D *C D *Z V *I

The above items represent five pitfalls to avoid in the construction of foils for multiple choice items.

1. Specific determiners. The use of absolutes ("never," "always," "totally," and "completely") suggests to students that these options are not correct answers. Students should be demonstrating their learning rather than their cleverness at deciphering answers to poorly constructed questions.
2. Cognates. A clue can give away the answer, often in the form of an option which resembles some word or phrase in the item stem.
3. Silly or ridiculous foils. This form of humor may provide comic relief during a test, but as a technique, it also may give the answer away.
4. Equivalent foils. If only one answer is allowed, and two options are equally correct, than neither can be the "right answer." This logic is often practiced by students.
5. Longest options are often the correct options. Poorly written items often have brief foils, thus increasing students' chances to make a lucky guess.

Other Deficiencies in Item Writing

The use of technical language and long passages ending with a question creates laborious reading, difficulty for weak readers. It also tries

students' patience, creates anxiety, and reduces motivation. Such questions take up too much testing time. To illustrate:

Fran is a high school student who wants to get good grades but has trouble studying due to distractions such as watching television, telephone calls from friends, and writing notes to her boyfriend. As a counselor, which of the following procedures for developing and maintaining motivation would you recommend to Fran?

- a. Interfere as little as possible, but provide learning experiences when requested.
 - b. Frequently remind Fran of the eventual value of what she is to learn.
 - c. Tell the parents not to interfere--Fran is probably engaging in nonstudy behavior to irritate them.
 - d. Pay Fran a specified amount for every hour of uninterrupted study time.
-

The multiple-multiple choice item can also be confusing and, hence, time-consuming.

When Flash arrived in Mongo, he

- I. went straight to the Palace.
- II. discovered that the Clay People had kidnapped Dale.
- III. found out that Doctor Zarkoff was being held in Ming's Palace.
- IV. was told that Prince Barron had been bickering with King Vulcan again.

- a. I and II
 - b. II and III
 - c. I and IV
 - d. II, III, and IV
-

This item is far too long. In addition, the clever student can figure out the best answer by noting the "II" is most often repeated, and therefore, option "c" is probably not correct. If the student knows that "I" is not correct, then "d" must be the correct answer.

Some item writers use tricks to lead students to select foils, a practice that is NOT recommended. The intent of the item is to measure achievement, not cleverness. Items which are direct and simply phrased yield the most valid information about student achievement. Tricks upset many students, consequently reducing the overall precision of the achievement measure.

Questions involving opinions, values, and attitudes should not be included in an achievement test. It is difficult to treat opinions as facts and test accordingly. If an opinion is to be tested, however, it should be qualified as in the statement below.

According to lectures, the best source of protein is:

Finally, options should be presented in random order. Item writers can slip into a pattern: 1. A 2. B 3. C 4. D; or spell words with options: 1. B 2. A 3. D. Students who have not learned, look desperately for clues.

True-False Tests

True-false questions, a form of multiple choice, are often useful. A greater number of true-false questions can be asked in an hour than any other type of question, making possible a wider range of content sampling. These tests are easier for many teachers to write than multiple choice, creating better efficiency. Although achievement in higher cognitive areas also can be measured with true-false questions, it is difficult. The primary disadvantage of the true-false item is its high guess level. A student who guesses randomly will get about 50 percent of the items correct. This problem can be alleviated, however, if one employs a great many items and is careful in the interpretation of test results. For example, a score of 54 percent on a true-false test would indicate a low level of achievement, with 85 percent a moderate level. By contrast, 54 percent on a multiple choice test would indicate moderate achievement and 85 percent would be fairly high.

Some tips for writing true-false items are listed below:

1. Deal with statements that are dichotomous in nature; avoid items having shades of meaning or degrees of comparison.

Good: The planet with two moons revolving around it is Mars.

Bad! One of the nicest things about Slobbovia is its climate.

The first item is clearly true, there is no conjecture. The second item is subject to some criteria for judging "niceness." It is a comparative judgment not well suited to true-false questions.

2. Avoid the use of negatives or double negatives. It has been shown that using negatives takes up more testing time and makes the test more difficult than the resulting information is worth.

Bad: You should not pour hydrochloric acid into water.
Good: A dangerous reaction will occur when hydrochloric acid is poured into water.

The careless, but knowledgeable, may omit reading "not" and answer incorrectly. The second phrasing of the question avoids the problem.

3. Avoid wordy sentences. Lengthy sentences are a test of the student's patience, concentration, and reading ability. Such questions are seldom effective measures.
-

Bad: A coding system may be defined as a set of contingently related, nonspecific categories which are not readily identifiable to the audience for which the system was intended but which are identifiable to the creators of the system.

Good: One example of a coding system is Bartlett's memory schemata.

In both cases, a coding system as it is related to thinking needs to be defined. While the first example is technically accurate, it is wordy, convoluted, and difficult to follow. The second example is accurate, brief, and easy to read.

4. All true-false items should contain single ideas. Trying to squeeze too many ideas into one question is ineffective.
-

Good: All squares have interior angles which sum to 360 degrees.

Bad: Squares, rectangles, and trapezoids have interior angles which sum to 360 degrees, and all quadrilaterals have four sides.

In the first example one concept is tested. In the second, at least three concepts are tested. The second example could be rewritten to produce three legitimate true-false items of merit:

- a. Squares, rectangles, and trapezoids are the three members of the quadrilateral family.
- b. All quadrilaterals have interior angles which sum to 360 degrees.

c. All quadrilaterals have four sides.

An excellent resource on writing true-false items is Ebel's Essentials of Educational Measurement, 1972 (see Appendix A).

Matching Items

When material is subject to more than four or five options and several items stems, matching items may be most efficient. In the following example, students are to identify which planets in the solar system have certain characteristics.

-
- | | |
|--|------------------|
| _____ 1. The speediest orbiting planet | a. Mercury |
| _____ 2. Has the densest atmosphere | b. Venus |
| _____ 3. Has traces of oxygen | c. Earth |
| _____ 4. Is a satellite | d. Mars |
| _____ 5. Is considered a moon | e. Jupiter |
| _____ 6. Has the most moons | f. Saturn |
| _____ 7. Has two moons | g. Uranus |
| | h. Pluto |
| | i. Neptune |
| | j. None of these |
| | k. All of these |
-

True-false and matching items, as variations of multiple choice, are effective for certain situations. The suggestions given earlier for multiple choice items would apply as well to these types of items.

Item Forms

One way to write multiple choice questions is to follow a model. Many test items can be created by simply inserting new words or elements into the model.

Consider for example, the math question:

What is the value of "X" in the equation $X + 4 = 10$?

- | | |
|-------|-------|
| a. 14 | c. 10 |
| *b. 6 | d. 4 |
-

This item serves as a model, or "form," for a whole "family" of items. By replacing the number +4 with any integer between -9 and +9, and replacing the number 10 with any number from 0 to 19, nearly four hundred questions can be created that measure the same domain of knowledge.

Another example:

Which of the following sentences has an INCORRECT use of the verb?

- a. Tomorrow, Mildred will be on vacation.
 - b. Yesterday, Melvin was tired.
 - c. Today, Fern weren't with it.
 - d. This morning, Jean wasn't here.
-

Sentences from a reading text could be identified and arranged into four-option sets. The verb tense in one of the options would be changed to provide an appropriate option.

The use of item forms is somewhat limited, for the present, to topics which are easily described (such as mathematics, spelling, English usage). However, the item form is one way to quickly create a great many selected response items.

Self-Quiz

Select the best answer. Write the letter corresponding to your choice in the space provided at the left.

- ____ 1. Which of the following is typically the most efficient form for the testing of knowledge for large groups of students?
 - a. product assessment
 - b. performance testing
 - c. selected response
 - d. constructed response

- ____ 2. The most important reason for using the multiple choice item is because of its
 - a. potential for better sampling of content
 - b. ease of scoring
 - c. flexibility in measuring behavior
 - d. objectivity

3. Option is to foil as
 a. answer key is to raw score
 b. objective test is to essay test
 c. possibility is to mistake
 d. multiple choice is to true-false
 e. objective is to subjective
4. Which of the following is not true of an objective item (as compared with an essay item)?
 a. has high scoring consistency from scorer to scorer
 b. can be scored quickly
 c. can be prepared quickly
 d. free of factors of skill in expression and penmanship
 e. free from opportunities for bluffing

For each of the following questions write either "+" for true or "0" for false in the space provided at the left.

5. The stem of a multiple choice item should state or clearly imply a specific direct question.
6. The stem of a multiple choice item should be limited to a single sentence or sentence fragment.
7. One option in each multiple choice item should be so absurd that it would be chosen only by a student who is guessing blindly.
8. Well-constructed true-false items are no more subject to guessing than are well-constructed four-choice multiple choice items.
9. It is often difficult and seldom advantageous to make all of the responses to a multiple choice item parallel in point of view, grammatical structure, or general appearance.
10. Multiple choice items having negatively stated stems (with the word not playing a crucial role) tend to be better items.
11. A well-written item can cause clever students to find the correct answer by a process of elimination of incorrect answers.
12. Good true-false items express single, not multiple ideas.

OUR ANSWERS

12. +	8. 0	4. C
11. 0	7. 0	3. C
10. 0	6. +	2. A
9. 0	5. +	1. C

CHAPTER III: CLASSROOM TEST CONSTRUCTION

Constructed Response Items

A test item which requires the student to create a response rather than to select one, is called a constructed response item. Math computation items are often in this category as are the more complex story problems frequently used in more advanced math and science content areas. The counterpart from other content areas is the completion question. The essay test is another type of constructed response test. Items are relatively easy to write, but often difficult to score due to the variety of correct responses available to students.

In many respects the essay test is the most misused of all constructed response types. Many teachers use it because it involves the student in written composition, not realizing the impact on measurement validity. The essay question should not be used to test knowledge of a given topic and to measure writing skill simultaneously. A student may know the answer, but the ability to communicate in writing may be lacking. The item writer must first decide whether writing skill or knowledge is to be measured. A short or extended answer essay may provide useful information about student knowledge of a subject; writing skills are more effectively measured by means of rating scales or checklists.

When Should the Essay Test Be Used?

Classroom achievement tests, when they are successful, accurately measure specific learning outcomes. As shown in earlier chapters, selected response tests hold the most potential for adequate sampling of content. The constructed response test is often limited to a narrower range of content.

A constructed response test must, of course, be used if the instructional outcome to be measured requires a written response. Consider the following example:

The student will describe in writing at least four features of the topography of Mars.

To test achievement of this objective, an essay question would ask the student to describe in writing four features of the topography of Mars.

The advantages and disadvantages of an essay test format are listed below:

Advantages

- + comparatively easy to prepare
- + lends itself easily to measuring higher level cognitive knowledge
- + may yield additional insights (or measures) of other learning

Disadvantages

- hard to score even when following the recommended procedures
- subject to a number of biases in scoring
- does not offer as good a potential for adequate coverage of a content domain

The extended answer essay is sometimes selected to measure the degree to which a student can select, organize, and synthesize material into a cohesive response. However, the extended answer essay is probably the least efficient in measuring school achievement because of the narrow range of content sampled, the confounding effect of writing ability, and the lack of precision in scoring.

The short answer essay is probably the most effective of the constructed response items; more questions can be asked in a given time period, as compared to the extended answer essay, and consequently more learning outcomes can be measured.

Writing a Short Answer Question

The following steps should be followed in writing a short answer essay question:

1. Determine in advance the concept or objective (learning outcome) to be tested.
2. Paraphrase the familiar material. Do not use verbatim material.
3. Use such verbs as "compare, illustrate, give examples of," so the student understands the nature of the task.
4. Make the questions as unambiguous and clear as possible.
5. Sample the content as fairly as possible. Do not overload the test with items from one particular area unless students have been prepared in this respect.

Perhaps the greatest difficulty in writing short answer essay questions is indicating the desired extent of the answer. Test writers often leave too much open to interpretation. For example:

Bad: How did Polk become President?

Better: Describe factors which contributed to Polk becoming President.

Best: List and briefly describe four major factors in Polk's life which contributed importantly to his becoming President.

The best short answer essay questions are detailed and clear. Less detailed questions lead to answers which, though still credible, may vary greatly from the model response. A well-written short answer essay question helps avoid the "creative" answer that is marginally acceptable.

Test Length and Format

Test accuracy by far is the most overriding concern to the test writer; learning outcomes must be fairly sampled to achieve accuracy. There are several options in essay testing in a fixed time period. With 45 minutes to test, ten short answer essay questions should provide enough sampling for accuracy. A few extended answer essay questions would not sample enough content. Avoid asking students to choose ten of twelve short answer essay questions, since this reduces the consistency of measurement from one student to another. (It is hard to write questions of "equal" difficulty.)

Scoring the Response

While the constructed response exam may be relatively easy to prepare, it is more difficult to score. The following suggestions can help realize effective and reliable scoring.

- Assign points to each question when writing the test. Give credit for the extent to which the student's answer fulfills the requirements of the model answer.
- Wherever possible, keep scoring anonymous. If not possible, keep in mind that knowing who wrote the answer can prejudice the scoring procedure.
- Prepare model answers for each question and use these answers in scoring. Sharing the model answers with the students after the test helps make the test a teaching tool as well as a measure.

- Try to keep poor grammar, penmanship, or spelling from biasing your judgment of the knowledge elicited by the item. If these skills are to be rated, they should be done so separately.
- If each item has a different weighting, be certain students understand this on beginning the test.
- Score one question at a time for all students, then proceed to the next question. This approach helps maintain consistency in scoring.
- Write comments on each student's paper. Offer advice, criticism, praise, suggestions. Students appreciate such feedback.
- If there are several sets of paper, shuffle them to avoid the tendency of downgrading earlier or later papers. Take rests and review the model answers to prevent fatigue and altering standards.

While selected response tests are usually the more efficient, constructed response tests can also be excellent measures when the class is small or the cognitive level of knowledge to be tested is high. (The essay format is more often used in college and graduate school courses.) If the short answer essay test is carefully written and scored, it can be a reasonably effective measurement tool.

Self-Quiz

Select the best answer. Write the letter corresponding to your choice in the space provided at the left.

- ____ 1. The major similarity between selected and constructed response items is that
 - a. both yield high reliability estimates
 - b. both are efficient with respect to administration and scoring
 - c. neither is completely valid
 - d. both are measures of school achievement

- ____ 2. One of the major difficulties with the essay item is that it
 - a. fails to obtain responses that differentiate among examinees
 - b. often fails to set uniform tasks for all examinees
 - c. seldom samples what is to be tested
 - d. suffers from the fact that examinees attempt to guess the correct response

- ____ 3. In the scoring of essay examinations, all of the following are considered desirable practices except to
- reduce the mark for poor spelling or penmanship
 - prepare a scoring key and standard in advance
 - remove or cover pupils' names from the papers
 - make individual comments on each student's paper
- ____ 4. The greatest advantage of short answer essay tests over selected response tests is
- the ease with which the test items can be constructed
 - the ease and accuracy with which such tests can be standardized
 - the ease with which the test results are interpreted
 - the better sampling of content

For each of the following questions, write either "T" for true or "F" for false in the space provided at the left.

- ____ 5. The best procedure in scoring an essay test is to read all questions on one student's paper before starting to read a second student's paper.
- ____ 6. A one-hour essay test composed of three questions requiring extended answers is likely to be more accurate than a one-hour essay test composed of twelve questions permitting much shorter answers.
- ____ 7. To make essay test scores objective in meaning is to defeat the purpose of essay testing.
- ____ 8. Different types of test items (essay, short answer, true-false, multiple choice, etc.) must be used to test different levels of cognitive behavior.

OUR ANSWERS

1. D 2. B 3. A 4. A 5. F 6. F 7. F 8. F

CHAPTER IV: CLASSROOM TEST CONSTRUCTION

Measuring Skills

For several decades the paper-and-pencil test has been the primary vehicle for measuring educational growth; it is only recently that a great deal of effort has gone into developing performance-based tests. The essence of the performance-based test is that if a performance or skill is to be measured, there should be direct, not indirect, means of measurement. For example, certain types of knowledge are essential to driving an automobile, yet having this knowledge in no way guarantees that one will be a skillful driver. Measurement of driving skill should involve actual driving performance tasks rather than knowledge-based paper-and-pencil tests. The former is a direct measure, the latter an indirect measure from which only weak inferences about driving skill can be drawn. To further illustrate this point, the most direct way to determine students' skill in doing laboratory experiments is to measure how well they perform a series of tasks essential to good laboratory technique. Determining knowledge of laboratory procedure does not yield direct information about lab skills.

The term "skill" is generally applied to behavior which involves specific processes which may or may not result in products. These processes and/or products have qualities or characteristics which can be directly observed. Judging the quality of these observable characteristics is what is meant by measuring skills. Learning outcomes of this kind are best measured by means of performance-based tests. In fact, the graduation competencies required by Oregon's new minimum standards imply many outcomes of this sort and necessitate the use of appropriate measurement techniques. Those described in this chapter include: observation, rating scales, and checklists.

In observation, one notes the presence or absence of an observable behavior; for example, a child tying a shoelace, putting things away, stacking blocks, or completing a puzzle. These are simple behaviors, many in the psychomotor domain.

Rating scales are numerical descriptions of behavior. The observer views a performance or product and records a numerical judgment on a rating scale. For example, when listening to an oral reading of Coleridge's "Rime of the Ancient Mariner," the teacher may rate the student's inflection (vocal variety) on a scale from one to five; one representing little inflection and five representing excellent inflection.

Checklists resemble observation, except a checklist focuses attention on sequences of related behaviors. A sequence might include assembling equipment or performing a series of tasks.

Observation

There are many instances where simple observation can be used to indicate whether or not an objective has been achieved; for example:

The student can:

1. dress self;
 2. run 200 yards in less than one minute;
 3. perform orally all multiplications involving one-digit numbers and two factors;
 4. read any paragraph orally with no more than two errors;
 5. spell all words on the Dolch reading list without error.
-

In performance-based tests of these tasks, the teacher notes whether or not the behavior has been demonstrated; there is no inference of knowledge beyond that which is stated. The reading example calls only for reading, not comprehension; the spelling example refers only to the specified list, not to other words on other lists. Therefore, the desired outcomes can be directly observed.

Observation is the simplest, most direct measurement technique known. It should be used whenever appropriate because of its high degree of reliability.

Rating Scales

If the quality or degree of skill achievement is to be measured, the rating scale is useful because it is accurate and efficient.

Advantages

- + simple to use
- + easy to interpret
- + requires test-maker to clearly define that which is being measured

Disadvantages

- subject to lack of agreement among raters
- may be time-consuming to administer
- usually involves inferences

While raters tend to differ in judgments, they can learn to increase precision in rating and hence, agreement. Identifying what is to be rated is perhaps the most important task in developing a rating scale. Simply creating scales to rate something as undefined as "creativity" or "motivation" is a dangerous practice; many interpretations can arise when a rater is examining a related performance or product.

The rating of traits usually calls for the exercise of judgment on the part of the observer. How well can a student adjust scientific equipment? How effective are the techniques used by a teacher to reinforce a poorly achieving student when homework is completed? These questions call for subjective judgments regarding a performance or product.

Four steps are recommended in constructing a good rating scale:¹

1. Describe clearly the trait(s) to be rated in the performance or product. For example, when rating students in a woodworking class on safe handling of tools, the word "safe" is not enough. A more adequate description might refer to using specific equipment, cleaning up after class, and following rules.
2. Create a scale for each characteristic. One of the most popular and useful types of scales is the graphic scale. For example:

How often do students clean up adequately?

☐

frequently

☐

about half
the time

☐

seldom

Here the rater simply marks the category that best fits the performance.

¹The following material has been adopted from Tenbrink (1974). For a more thorough treatment of techniques to measure performances and products, see Tenbrink, pages 273-293.

In other instances, a verbal descriptive rating scale is used:

How well are colors used in the batik?

a. very well b. well c. average d. poorly e. very poorly

Table 3 presents a variety of response options, each based on a five-point or a three-point scale. A five-point scale is generally preferred, although three-point scales, seven-point scales, and even ten-point scales have been used effectively. Three-point scales usually do not capture the entire range of a trait while the larger point scales call for too fine a discrimination by the rater.

TABLE 3

Examples of Various Types of Rating Scales

SIMPLE NUMERICAL

Rate the following using this scale:

1 = excellent 2 = good 3 = average 4 = poor 5 = very poor

1 attention span in class
3 able to follow directions

SIMPLE GRAPHIC SCALES

Our Textbook: ☐ like ☐ indifferent ☐ dislike

E How much has the performance changed since the last time?

A. much better B. better C. about the same D. worse E. much worse

B In terms of originality, how would you rate the essay you have read?

A. high B. average C. low

- A How well did the student read the passage?
A. very well B. well C. as well as most (average) D. poor E. very poorly
- C How often did the student correctly use the workbook?
A. very often B. often C. sometimes D. seldom E. never
- B To what degree does the exhibit contain detail?
A. very much B. much C. some D. little E. very little
- D The performance lacked volume.
A. strongly agree B. agree C. neither agree nor disagree D. disagree E. strongly disagree
-

DESCRIPTIVE SCALE

☐	☐	☐	☐
not meeting the requirements	fair but needs improving	satisfactory	doing good work
☐	☐		☐
excellent job	develops paragraph quite well with clear-cut topic sentence		about average development of paragraph
☒			
does not develop paragraph well, frequently lacking topic sentence and adequate definitions			

Factors Affecting the Reliability of Ratings

Once a trait is clearly described and a scale established, the rating should be a valid reflection of student achievement. However, the following factors may interfere with the usefulness of the rating scale:

- Lack of interest; if the rater is bored with the task, the results will reflect this.
- Personal bias interferes with judgment and results are liable to be distorted.
- Extreme options (e.g., never, always) seldom provide useful information. If a five-point scale contains two absolutes they will rarely be selected.
- Lack of clarity about what is being rated will cause erratic results.
- Generosity; there is a tendency to overrate when scales are used.
- Halo; there is a tendency to give a global rating to a person and make ratings of sub-tasks correspond with this rating. Again, this distorts the results.
- Interaction errors. When a panel of judges makes independent ratings, high and low scores can be omitted and the remainder averaged.

Checklists

As noted earlier, the measurement of many skills does not require a rating. The presence or absence of a desired outcome could be observed or, if there is a need to measure whether or not steps leading to an outcome have been achieved, a checklist would be more appropriate.

EXAMPLE:

Brewing a pot of coffee:²

- | | |
|---|--|
| ☐ disconnect coffee pot | ☐ fill basket with coffee |
| ☐ disassemble coffee pot | ☐ reconnect coffee pot |
| ☐ clean pot | ☐ set dial on pot |
| ☐ fill pot with water | ☐ check to see if perking |

²Adapted from Mager's Measuring Instructional Intent, 1974, p. 11.

EXERCISE

Ten potential checklist items are provided below. Place an "X" in the space if the item is visible and easily demonstrated. The student:

- ___ 1. reads a passage without making the following types of errors . . .
- ___ 2. correctly punctuates contractions, words requiring apostrophes, and plural possessives
- ___ 3. is happy
- ___ 4. is independent
- ___ 5. performs each of twelve simple life skills
- ___ 6. puts all equipment away in correct places after experiment
- ___ 7. can locate each of the following types of symbols on any standard map using the legend . . .
- ___ 8. is adjusted to classroom
- ___ 9. follows five steps in using a platform balance
- ___ 10. has mastered each of the five word attack skills in this unit

If you checked 1, 2, 5, 6, 7, 9, and 10, you have correctly identified outcomes appropriate for checklists.

Constructing checklists is similar to making rating scales. Be sure you:

- 1. Describe the product or performance adequately.
- 2. List the behaviors to be observed in correct sequence.
- 3. Note errors which may occur in the performance.
- 4. Give clear directions about how to use the checklist.

The checklist is most applicable to simple, observable performances. It is easy to develop and use, and the reliability can be high. Checklists are becoming increasingly popular. Some school districts are even reporting student progress to parents in terms of what students can or cannot do, rather than using the traditional report card.

SELF-QUIZ

Mark "O" if appropriate for observation
"R" if appropriate for rating scales
"C" if appropriate for checklists

- ___ 1. How well can Jascha play the violin?
 - ___ 2. Has Andrea been in school everyday this term?
 - ___ 3. Has George completed all seven steps correctly in the experiment?
 - ___ 4. Can Gary follow all 12 steps in correctly reassembling a simple motor?
 - ___ 5. How much improvement in soccer has Pele shown in P.E. this year?
-

Select the best answer. Write the letter corresponding to your choice in the space provided at the left.

- ___ 6. From the standpoint of construction, which technique is most liable to offer the highest efficiency?
 - a. rating scales
 - b. checklists
 - c. observation
 - d. none of these
- ___ 7. Which requires judgment by the observer?
 - a. rating scales
 - b. observation
 - c. checklists
 - d. none of these
- ___ 8. A checklist must
 - a. be administered by more than one person
 - b. refer to intended behavior
 - c. lead to a numerical result
 - d. have a logical sequence of behaviors to be checked
- ___ 9. Observation is to checklist as
 - a. direct is to indirect
 - b. simple behavior is to sequential behavior
 - c. inference is to induction
 - d. deduction is to itemizing

Mark "+" if a good practice and "0" if not, in the use of observation scales, rating scales, or checklists.

- ____ 10. First, Mr. Ree makes an overall rating of his drama class, then uses that rating for each aspect of student performance.
- ____ 11. Marion Polk believes that all history exhibits should be judged by at least three persons.
- ____ 12. Counselor Guy Nice, when judging for science fairs, insists that "good intentions" be considered in assigning ratings.
- ____ 13. Polly Gohn uses a checklist for each solution to a proof to see if all steps are followed.
- ____ 14. Atlas Shrug, the P.E. teacher, uses a rating scale to assess the degree to which he has achieved his goals each term.

OUR ANSWERS

+	14.	7.	A
+	13.	6.	C
0	12.	5.	R
+	11.	4.	C
0	10.	3.	C
8	9.	2.	O
D	8.	1.	R

CHAPTER V: MEASURING ATTITUDES

Attitude can be described in several ways: it is a tendency to act in certain ways under certain conditions; it is also viewed as a tendency to react emotionally, either positively or negatively. We infer the attitudes of people from what they say or do. In this chapter, we will look at student attitudes about school, subject matter, policies, practices, environment, and other parts of their educational program.

Teachers are interested in attitude for two reasons. Students' perceptions about school have much to do with what they learn; subject matter that is disliked is not easily absorbed. Educators have begun to look at positive attitudes as an important school outcome. Since students spend a good portion of their childhood and adolescence in schools, the school environment should be seen by them as generally constructive.

Approaches to Measuring Attitudes

Two methods are generally used to measure student attitudes: self-report by the student and observation of the student by others. With the self-report, the student is asked to tell how he or she feels. Responses can be elicited through interviews or questionnaires of various types. The observation method requires the teacher or some other person to observe the student's behavior. With the anecdotal approach, essay reports are usually employed to describe student behavior in a particular incident.

Two self-report methods are used to learn how students feel: they can be interviewed, usually one at a time, or they can complete a questionnaire. Interviews may not be very efficient in terms of time required or costs. A more serious drawback is that interviewers must be carefully trained in interview techniques if reliable information is to be obtained.

The questionnaire is more commonly used. The teacher asks questions of students as a group, and they write their responses. Two assumptions are necessary: the questions will mean the same thing to all respondents, and all respondents will answer the questions honestly.

Advantages

- + can be done anonymously
- + a simple and direct way to gain information of a systematic and quantifiable nature
- + most reliable when raters are qualified

Advantages (Continued)

- + most valid in instances where objects to be rated (e.g., textbooks, classroom topics) are well-defined
- + can be given to groups of students instead of one student at a time

Disadvantages

- can be difficult to construct and analyze
- can be time-consuming to analyze
- may be subject to dishonesty by students
- cannot identify how particular individuals feel because of the anonymity

While a great deal of time is spent constructing useful items for the self-report questionnaire, the instrument can be used in a variety of situations to obtain individual or group information.

The self-report scales most commonly used are the rating scale and pair comparisons. The rating scale is the most direct. The precision of a rating scale tends to increase with the number of scale points provided; the most practical number of scale points seems to be five. An odd number of scale points (e.g., five or seven) allows for a "neutral" point which many respondents find useful. Without a neutral point, respondents are "forced" to respond in a definite direction, which may misrepresent their views to some extent. For example:

How do you feel when it is time to leave school?

a. very happy ☒ b. happy c. neutral d. sad e. very sad

One successful variation of the rating scale for elementary school students is the use of symbols for the response options. The options are reduced to three and represented in the following manner:

How do you feel when it is time to leave school?

a. ☒  b. ☐  c. ☐ 

Studies by the Teaching Research Division, Monmouth, Oregon, confirm this method as an effective means of ascertaining children's attitudes as early as the first grade.

Rating scale items can be used for a variety of purposes, such as assessing perceived learning effects, feelings, strength of feelings, or agreement-disagreement. Some examples:

-
1. What effect (e.g., on your learning) does (something) have on you?
a. very much ☒ b. much c. moderate d. little e. very little
 2. How do you feel about (something)?
a. very satisfied b. satisfied c. neutral ☒ d. dissatisfied e. very satisfied
 3. To what extent do you agree with (a statement of opinion)?
a. strongly agree ☒ b. agree c. neither agree nor disagree d. disagree e. strongly disagree
-

The pair comparison is effective when measuring the strength of student preferences between objects, activities, or subject matter areas. This technique helps establish an order of expressed attitudes. The example below seeks to determine preference of math versus reading.

Which would you rather do?

- | | | |
|---|----|---|
| ☐ a. read a book | OR | ☒ b. work math problems |
| ☒ a. do math homework | OR | ☐ b. do reading homework |
-

With either type of format, the following recommendations apply when constructing the instrument:

1. Use phrases that are simple and understandable to the students.
2. Phrase statements so that some are clearly negative and some are clearly positive, with approximately the same number of each.
3. Keep all items relevant; avoid making the questionnaire over-long.
4. In using the pair comparison technique, be sure that choices include all possible combinations.

Scoring and Analyzing Results

The rating scale is generally scored with number values being assigned to each of the options. If an item is reversed (that is, stated negatively instead of positively), the number scale is reversed. The attitude measure is the sum of the values of the responses. In the following example, student attitude toward physical education is measured.

Circle the best answer for you.

1. How do you feel when P.E. begins? (positively stated)

☒ a. happy	b. neutral	c. sad
3	2	1
2. How do you feel when you are in P.E.? (positively stated)

a. sad	b. neutral	☒ c. happy
1	2	3
3. How do you feel when P.E. is over? (negatively stated)

a. happy	b. neutral	☒ c. sad
1	2	3
4. How would you feel if you never had to go to P.E. again? (negatively stated)

☒ a. sad	b. neutral	c. happy
3	2	1

SCORE: 12 (The score of 12 is the highest possible; 4 would be the lowest possible.)

The pair comparison technique requires a different scoring formula. Consider the example below of a questionnaire administered to a teacher.

Mark the box indicating which activity you prefer:

- | | | |
|--|----|---|
| 1. ☒ going to school | OR | ☐ going to a faculty meeting |
| 2. ☒ attending workshops | OR | ☐ collecting milk and lunch money |
| 3. ☐ going to school | OR | ☒ attending workshops |
| 4. ☐ going to a faculty meeting | OR | ☒ collecting milk and lunch money |
| 5. ☒ attending workshops | OR | ☐ going to a faculty meeting |
| 6. ☒ going to school | OR | ☐ collecting milk and lunch money |

Responses are tallied and four attitude scales are constructed.

<u>Preference</u>	<u>Frequency</u>	<u>Score</u>
going to school	//	2
going to a faculty meeting		0
attending workshops	///	3
collecting milk and lunch money	/	1

For this simple attitude survey, the scale for each activity runs from zero to three. This teacher likes going to workshops, enjoys going to school next, and likes faculty meetings the least of the four activities. The pair comparison approach is particularly useful for revealing relative differences in attitude.

Observation Methods

As described earlier in the chapter on measuring skills, observation methods are direct, and the advantages in measuring attitudes are similar to those for measuring skills. The observation approach requires little time for construction or tabulation. Attitudes reflected by the general day-to-day behaviors of students can be observed; attention can also be focused on critical incidents and observations made of behaviors in stress situations.

General Observations

One way to classify behavior that reflects attitude is in terms of approach or avoidance. If a feeling toward something is positive, one tends to seek it; when feelings are negative, one tends to avoid it. The following observations illustrate approach and avoidance for one student in a school setting.

-
1. According to his mother, Jason spends most of his time each evening reading new books from his school book club.
 2. During recesses, Jason goes to the library to read.
 3. At P.E., Jason often says he doesn't feel well.
 4. After school most of the kids stay on the playground to play; Jason usually goes home.
 5. During free time each day, Jason goes to the library.
 6. Jason has ordered more new books from his school book club than any other student.
 7. A survey of attendance shows that Jason never misses morning sessions of school, and frequently misses afternoon sessions.
-

Approach and avoidance behaviors are easy to observe, systematic in nature, and reflect Jason's preferences. They are simple observations requiring no inference. The sum of these observations, however, may be used to draw inferences about Jason's attitudes toward reading and P.E.

Observation can also be used for group measures. Mager (1974, p. 89) offers some attitude measures that fall into this category, some of which are listed below.

-
1. Percentage of students completing the course.
 2. Number of students volunteering to transfer into someone else's class.
 3. Number of papers or projects longer than required.
 4. Number of assignments completed on time.
 5. Frequency of use of a particular learning center.

6. Number of students volunteering to stay after school to help.
7. Number of library books and other materials checked out on the subject.
8. Increase or decrease in school vandalism and petty theft.

These are a few of the many possible behavioral measures that teachers can use to estimate how a teaching strategy, program, new technique, materials, or policies may be affecting student attitudes.

Many of these measures are "unobtrusive" in that they do not interfere with student activities or class time, and students may not be aware that the measures are being taken. Information about students who are sensitive to being observed is less likely to be "faked" or biased when unobtrusive measures are used. While observation methods have the attractive quality of being simple and direct, they can be time-consuming and demanding on the teacher. For those wishing to further pursue attitude measurement along these lines, consult Mager's Developing Attitude Toward Learning (see Appendix A).

Critical Incidents

No matter how carefully the school activities are managed, every student experiences stress situations from time to time which we could call "critical incidents." Careful observation of these situations can often lead to very useful insights into student attitudes. The method differs from the general observation approach; behaviors are observed and interpreted within the context of the situation in which they occurred.

For example, general observation might indicate that Jeff loses his temper in class at least once a day. Knowledge of this may let the teacher know that Jeff needs to learn how to control his temper better. Careful observation of each incident may suggest some good ways to help him.

There are some rules to follow in recording useful observations of this kind:

- Care must be taken to record in sequence the facts of the situation separately from any interpretation (e.g., Jeff struck John in the face with his fist is fact; to say Jeff was angry when he did it is an interpretation).
- The conditions and behaviors which led up to the incident (the antecedents) must be carefully recalled and noted.
- The observer's (teacher's) responses to the situation must be recorded (again, facts only).

- The outcome of the situation must be factually recorded.
- Interpretations should be recorded separately and based on a thoughtful and logical analysis of this particular situation, its antecedents, observer (teacher) response, and the outcome.

Periodic review of accumulated descriptions may begin to reveal patterns of conditions under which Jeff loses his temper and conditions under which he is able to control it. A general format might look something like the following:

Description of Situation	Interpretative Comment
Antecedents:	:
Incident:	:
Teacher Response:	:
Outcome:	:

Caution is warranted. The critical incident method is of value only to the person using it. It is designed to help teachers separate the facts of a situation from the emotional impact it can have on them. Any actual recording of descriptions should be viewed only as the teacher's own personal notes for sorting out and creating conditions that can foster constructive student attitudes and behaviors. As soon as they have served this purpose, they should be destroyed.

Self-Quiz

Select the best answer. Write the letter corresponding to your choice in the blank provided at the left.

- ____ 1. Which one of the following best exemplifies an attitude?
 - a. choosing math over reading
 - b. losing your temper when disappointed
 - c. ignoring a best friend when unhappy
 - d. being happy-go-lucky
- ____ 2. Student attitudes are important to the teacher because they are
 - a. predictive of success in school
 - b. important school outcomes
 - c. both a and b
 - d. neither a nor b
- ____ 3. Which one of the following is a major disadvantage of a self-report?
 - a. lacks precision
 - b. tendency of some students to fake answers
 - c. time-consuming to construct the instrument
 - d. is group-administered
- ____ 4. A pair comparison technique is useful for
 - a. making comparisons between two objects
 - b. a self-report, when rating scale is inappropriate
 - c. finding the strength of a preference
 - d. determining the order of preferences for a set of objects
- ____ 5. Which of the following is generally false?
 - a. Self-reports are more desirable than interviews from the standpoint of efficiency.
 - b. A pair comparison is a more indirect measure than a rating scale.
 - c. Training of raters improves the precision of ratings.
 - d. Observations are typically very objective measures.
- ____ 6. Critical incident observations are well-done when
 - a. the record contains some impressionistic information
 - b. a complete description of a series of events and the student is done
 - c. psychological interpretations of actions are made
 - d. observed behavior of a student is described and interpreted from beginning to final outcome.

OUR ANSWERS

1. A 2. C 3. B 4. D 5. B 6. D

APPENDIX A

ANNOTATED BIBLIOGRAPHY

The following publications represent a sampling of recent and significant contributions to the technology of testing.

- Baker, E. L., and Popham, W. J. Expanding Dimensions of Instructional Objectives, Englewood Cliffs, Prentice-Hall, 1973. This brief text describes the role of objectives in education, the use of needs assessment in choosing goals, and some affective objectives. Writing tests to measure objectives is also discussed.
- Block, J. H. (Ed.) Mastery Learning: Theory and Practice. New York, Holt, Rinehart, and Winston, 1971. A thorough review of studies attesting to the benefits and deficits of mastery learning approaches to instruction.
- Bloom, B. S. Learning for Mastery. Evaluation Comment, 1968, 1. 1-12. One of the earliest and most cogent writings on mastery learning in American education.
- Dizney, H. Classroom Evaluation for Teachers, Dubuque, William C. Brown, 1971. A brief basic text on achievement testing which covers the fundamentals of testing.
- Ebel, R. Essentials of Educational Measurement, Englewood Cliffs, Prentice-Hall, 1972. A standard and widely used text in constructing and using teacher-made tests. The chapter on true-false testing is one of the best treatments of the subject.
- Gronlund, N. E. Improving, Marking, and Reporting in Classroom Instruction, Riverside, Macmillan, 1974. This is a very practical little booklet which contains information about grading practices, systems, and guidelines for collecting and using data to make criterion-referenced and norm-referenced decisions.
- Gronlund, N. E. Individualizing Classroom Instruction, Riverside, Macmillan, 1974. This booklet, one of a series by the author, describes the role of testing in individualized instruction. He describes several currently operating systems which are individualized.
- Gronlund, N. E. Preparing Criterion-Referenced Tests for Classroom Instruction, Riverside, Macmillan, 1973. This book is one of the few that attempts to describe how criterion-referenced tests are actually constructed and used.
- Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10. This is one of the most readable accounts of domain-referenced testing as it presently exists.

ANNOTATED BIBLIOGRAPHY, Continued

- Keller, F. S. Goodbye, teacher Journal of Applied Behavior Analysis, 1968, 1, 79-89. An account of Fred Keller's Personalized Student Instruction (PSI), one which motivated a movement in college instruction away from traditional ways toward competency-based approaches.
- Kibler, R., Barker, L. L., and Miles, D. T. Objectives for Instruction and Evaluation, Boston, Allyn and Bacon, 1974. This book provides the rationale for the use of behavioral objectives in systematic instruction. It also contains information on the selection, writing, and appropriate uses of objectives in teaching.
- Mager, R. F. Developing Attitude Toward Learning, Belmont, Fearon Inc., 1968. If ever a book were written for teachers that presents the reasons for being concerned about children's attitudes and how to measure it, this is the one. The book is both relevant and enjoyable to read.
- Mager, R. F. Goal Analysis, Belmont, Fearon Inc., 1972. In Mager's entertaining way, he describes appropriate methods to analyze the nature and demands of any goal. The applicability of this book to education is not apparent. But there are implications that need to be realized.
- Mager, R. F. Measuring Instructional Intent, Belmont, Fearon Inc., 1973. A common sense approach is taken by Mager in this book to drafting test items which directly reflect your instructional intent. Some important distinctions are drawn between norm- and criterion-referenced tests.
- Miller, H., Williams, R., and Haladyne, T. Beyond Facts: Objectively Measuring Higher Level Thinking, in press. The authors present and develop a system for writing multiple choice test questions that measure higher level thinking. Examples, exercises, and self-quizzes are used to develop the item writer's skills.
- Popham, W. J., and Baker, E. L. Establishing Instructional Goals, Englewood Cliffs, Prentice-Hall, 1970. This brief text describes the role of objectives in systematic instruction. It is replete with examples and exercises.
- Popham, W. J., and Baker, E. L. Planning an Instructional Sequence, Englewood Cliffs, Prentice-Hall, 1970. This book describes the way objectives are used in instruction. Like others in the series, it is rich with examples and exercises.
- Popham, W. J., and Baker, E. L. Systematic Instruction, Englewood Cliffs, Prentice-Hall, 1970. This book, from a series by the two authors, represents an overall look at modern instructional approaches which require the use of objectives and objective-based tests.

ANNOTATED BIBLIOGRAPHY, Continued

Popham, W. J., and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9. A more technical treatment on criterion-referenced testing, one of the earliest and most authoritative.

Sanders, N. Classroom Questions, Scranton, Harper and Row, 1966. This book is devoted to developing the idea that test items can be constructed to represent levels of Bloom's cognitive taxonomy. The book has many examples.

Southwest Regional Laboratory for Educational Research and Development, Educational Criterion Measures, Cincinnati, Van Nostrand, 1971. This short booklet is one of seven in a series called Instructional Product Development. It is useful in suggesting a wide variety of nonpaper-and-pencil measures of school behavior.

_____, Stating Educational Outcomes, Cincinnati, Van Nostrand, 1971. Another booklet in the Instructional Product Development series which describes how to effectively write or select objective. It comes complete with objectives, examples, and exercises.

TenBrink, T. D. Evaluation: A Practical Guide for Teachers, New York, McGraw-Hill, 1974. This book, a useful general guide for evaluation, contains a very concise and practical treatment of questionnaire development and sociometric instrument use. The book also deals with the problems of constructing teacher-made tests.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechres, T. L. Unobtrusive Measures, Chicago, Rand McNally, 1966. Novel methods for gaining information for a wide variety of purposes is discussed in this book in a most entertaining and interesting manner.

APPENDIX B

GLOSSARY OF TERMS

AFFECTIVE DOMAIN: One of three aspects of human behavior, dealing with attitudes, values, sentiments, feelings, personality, and other similar concepts. Cognitive behavior may be importantly related to affective behavior, but the concern in measuring affective behavior is that of ascertaining the degree of attitude, value, etc.

ATTITUDES: A tendency to react emotionally to an object in a positive, neutral, or negative way.

CHECKLISTS: A device for noting the presence or absence of behaviors which are sequentially related or conceptually organized. Used to study skills in the cognitive domain and aspects of affective behavior and psychomotor behavior.

CLASSROOM ACHIEVEMENT TEST: Any test expressly designed to measure the learning that has occurred as a direct result of classroom instruction.

COGNITIVE DOMAIN: One of three aspects of human behavior, dealing with intellectual activities.

CONSTRUCTED RESPONSE TEST: A test deliberately constructed so that the student must compose the answer. Used primarily to measure achievement of knowledge in the cognitive domain.

DOMAIN-REFERENCED TEST: Any test constructed by sampling from a collection (pool) of items related to a specific content domain.

EFFICIENCY: A test characteristic. A test is highly efficient if it is simple and inexpensive to plan, construct, administer, score, and report.

GROUP-REFERENCED: A manner in which achievement data may be referenced. That is, student performance is compared with the performance of others via group statistics such as the mean (average), median, or mode.

INSTRUCTIONAL OBJECTIVES: A statement of instructional intent which includes an action verb, conditions under which the student must perform, and a desirable level of performance.

OBSERVATION: A method for measuring simple types of skills, attitudes, or psychomotor behaviors.

PSYCHOMOTOR DOMAIN: One of three aspects of human behavior, dealing primarily with physical (psychomotor) performance. One should recognize that cognitive behavior is certainly related to psychomotor behavior. Measures of psychomotor behavior are primarily focused on the physical aspects rather than knowledge.

RATING SCALE: A device for developing numerical descriptions of skills, attitudes, or psychomotor behaviors. These numerical descriptions are essentially judgmental in nature.

RELIABILITY: A test characteristic. A test is reliable if it yields scores which are precise and consistent over time (i.e., stability, repeatability, and consistency of the measurement).

SELECTED RESPONSE TEST: A test deliberately constructed so the student must choose the correct answer from a set of options.

SKILLS: Attributes or characteristics inferred or directly observed through the consideration of performances or products. Observation, rating scales, or checklists are customarily used to measure skills.

TEST: A type of measurement where the conditions for the measurement are uniform for all examinees.

VALIDITY: A test characteristic. A test is valid if it measures what it is purported to measure.

APPENDIX C

A FEW ITEM SOURCES

Northwest Evaluation Association Item Bank
c/o Dr. Frederick V. Forster
PO Box 3107
Portland, OR 97208
(503) 234-3392

Education Commission of the States
National Assessment of Educational Progress (NAEP)
300 Lincoln Tower
1860 Lincoln Street
Denver, CO 80203

Instructional Objectives Exchange (IOX)
Box 24095, Department V
Los Angeles, CA 90024

Minnesota Educational Assessment
Minnesota Department of Education
Minneapolis, MN 55435

Michigan Department of Education
State Department of Education
Lansing, MI 48901

MEASURING PERFORMANCE: TEACHER MADE TESTS

YOUR VIEWS ARE IMPORTANT! After you read and examine this publication, please forward your comments to the publications staff of the Oregon Department of Education. If you would rather talk by telephone, call us at 378-4776. Or, for your convenience, this response form is provided.

PLEASE RESPOND so that your views can be considered as we plan future publications. Simply cut out the form, fold and mail it back to us. We want to hear from you!

Did you read this publication?

- ☐ Completely
- ☐ More than half
- ☐ Less than half
- ☐ Just skimmed

Does this publication fulfill its purpose as stated in the preface or introduction?

- ☐ Completely
- ☐ Partly
- ☐ Not at all

Did you find this publication useful in your work?

- ☐ Often
- ☐ Sometimes
- ☐ Seldom
- ☐ Never

Which section is most valuable? _____

What type of work do you do?

- ☐ Classroom teacher
- ☐ Consultant to classroom teachers
- ☐ School administrator
- ☐ Other _____

Would you recommend this publication to a colleague?

- ☐ Yes, without reservations
- ☐ Yes, with reservations
- ☐ No
- ☐ Other _____

Did you find the content to be stated clearly and accurately?

- ☐ Always yes
- ☐ In general, yes
- ☐ In general, no
- ☐ Always no
- ☐ Other _____

Were the contents presented in a convenient format?

- ☐ Very easy to use
- ☐ Fairly easy
- ☐ Fairly difficult
- ☐ Very difficult
- ☐ Other _____

Did you find this publication to be free of discrimination or biased content towards racial, ethnic cultural and religious groups, or in terms of sex stereotyping?

- ☐ Yes, without reservations
- ☐ Yes, with reservations
- ☐ No
- ☐ Other _____

What is your impression of the overall appearance of the publication (graphic art, style, type, etc.)?

- ☐ Excellent
- ☐ Good
- ☐ Fair
- ☐ Poor

When this publication is revised, what changes would you like to see made? _____

Additional comments. (Attach a sheet if you wish.)

Thanks!

Fold here and seal



BUSINESS REPLY MAIL

FIRST CLASS PERMIT NO. 199, SEC. 910, P. L. & B. SALEM, OREGON

Communications/Government Relations
Oregon Department of Education
942 Lancaster Drive NE
Salem, Oregon 97310

Fold here and seal